Published in History and Philosophy of the Life Sciences, 2010

Documenting the Emergence of Bio-Ontologies:

Or, Why Researching Bioinformatics Requires HPSSB

Sabina Leonelli

ESRC Centre for Genomics in Society

University of Exeter

Byrne House

St Germans Road

EX4 4PJ Exeter, UK

Tel. 0044 1392 269137

s.leonelli@exeter.ac.uk

Short title for running head:

Documenting the Emergence of Bio-Ontologies

Abstract (197 w)

This paper reflects on the analytic challenges emerging from the study of bioinformatic tools recently created to store and disseminate biological data, such as databases, repositories and bio-ontologies. I focus my discussion on the Gene Ontology, a term that defines three entities at once: a *classification system* facilitating the distribution and use of genomic data as evidence towards new insights; an *expert community* specialised in the curation of those data; and a *scientific institution* promoting the use of this tool among experimental biologists. These three dimensions of the Gene Ontology can be clearly distinguished analytically, but are tightly intertwined in practice. I suggest that this is true of all bioinformatic tools: they need to be understood simultaneously as epistemic, social and institutional entities, since they shape the knowledge extracted from data and at the same time regulate the organisation, development and communication of research. This viewpoint has one important implication for the methodologies used to study these tools, that is the need to integrate historical, philosophical and sociological approaches. I illustrate this claim through examples of misunderstandings that may result from a narrowly disciplinary study of the Gene Ontology, as I experienced them in my own research.

> The circuits traced out by these contemporary economies of vitality are thus conceptual, commercial, ethical, and spatial. These spaces range from the atomic, the molecular, the cellular, the organic, the spaces of practices (laboratories, clinics, consulting rooms, factories), of cities and their economies (Shangai, Munbai, the Cape), of nations and their regulatory frameworks and economic strategies, and of virtual spaces of the Internet that ensure the immediate availability at any point in the world of the totality of data on the genome. [..] The production of the exploitable knowledge of vitality today thus involves multiple transnational circuits to mobilize and associate material artefacts, tissues, cell lines, reagents, DNA sequences, techniques, researchers, funding, production, and marketing (Rose 2007, p.38).

Introduction

Within science studies, increasing attention is being dedicated to data sharing practices as a key aspect of the "economies of vitality" of which Rose speaks. Research on organisms involves increasingly complex networks of funding and exchange on a global scale. The creation of cyberinfrastructures, broadly defined as digital tools facilitating cooperation through the Internet, epitomises this state of affairs: billions are invested to ensure that data produced within diverse areas of inquiry are circulated as widely and efficiently as possible, in the hope that open access to data will enhance opportunities for new insights.

The 21st century has brought immense technological advances in the production of genomic data. Sequencing is now an automated activity taking no more than a few hours; collecting data on gene expression can also be done automatically, resulting in billions of data-points per day. This level of automation means that the activity of data collecting has never been as disjointed from activities of explicit theory-building. Contrary to data resulting from experiments associated to the testing or the production of hypotheses, automatically produced data are not explicitly associated to specific

hypotheses about the phenomena that they are documenting, other than to the theoretical assumptions used to build the instruments through which data were produced. More than any other source of evidence, automatically produced data require great interpretive efforts to determine what they can be evidence *for*. The scientific focus is thus shifting from efforts to produce data, characteristic of late 20th century biology, to efforts to exploit these data as evidence towards new claims.

Bioinformatic tools such as databases, bio-ontologies and repositories, as well as the scientists ('curators') in charge of developing them, are among the protagonists of this effort in contemporary life science. Funding bodies and researchers agree that the way in which data are interpreted depends not only on the skills and interests of the scientists using them as evidence, but also on how they are classified and formatted for dissemination. Accordingly, whoever is in charge of classification practices holds a considerable amount of power over the ways in which research will be conducted in the future.[1] This opens an interesting set of questions for the science studies community: what is the relation between data sharing and data use? How are data disseminated, and to what effect? What is the impact of cyberinfrastructure on the content, organisation and future of research? And in which sense precisely are these practices 'new', if at all?

This paper reflects on the difficulties I experienced in my own attempts to tackle these questions. My main claim is that digital tools for data sharing play multiple and interconnected roles within contemporary biology, and that this multiplicity can only be captured through an interdisciplinary approach including historical, philosophical and sociological perspectives. To give substance to this claim, I discuss some problems that derive from studying bioinformatics from a narrowly disciplinary viewpoint, and how a multiple approach may help to tackle those problems. While I am aware that my general argument is well-rehearsed within science studies of biology, I hope that my discussion can interest scholars dealing with the complex methodological issues raised by the study of bioinformatics.

I focus on the case of the Gene Ontology (GO), a digital tool devised at the turn of the millennium to provide free online access to genomic data obtained on the most popular model organisms. GO works by enabling researchers to search through existing model organism databases. It provides umbrella

[1] For an example of how the introduction of bioinformatics into classification practices shifted the balance of power in phylogeny, see Suàrez-Diàz and Anaya-Muñoz (2008).

categories through which data hosted in these databases can be retrieved on the basis of the users' research interests, regardless of the experimental context (including the specific organism) in which they were originally gathered. Focusing specifically on gene products, GO has become a very successful system for disseminating data, an exemplar for the development of other bio-ontologies in almost every data-rich realm of the biological and biomedical sciences.[2] My discussion is based on my research experiences in investigating GO between 2004 and 2008, and it starts by examining some interpretive pitfalls connected to a purely philosophical approach. I then discuss ways in which an historical and a sociological perspective may help to avoid those pitfalls.[3] I conclude by reflecting on methodological and professional concerns arising from the integrated HPSSB approach that I support.

1. Philosophical perspective: GO as a classification system

I first came across GO when visiting the Carnegie Institute of Plant Biology in Stanford University, California, in July 2004. I was investigating the development of The Arabidopsis Information Resource (TAIR), a database collecting data on the model plant *Arabidopsis thaliana*, and the director of TAIR invited me to sit in a GO 'content meeting'. This was a meeting organised to discuss the definitions of three terms in GO: "metabolism", "pathogenesis" and "cell cycle". The twenty-four participants included the curators of GO, the curators of prominent model organism databases that use GO to classify their data, and a few leading experimental biologists with expertise in the terms in question. What took place during that meeting captured immediately my philosophical imagination. Here was a high-profile group of biologists with diverse interests and skills, gathering to debate not only the biological meaning of the term pathogenesis, but also the methodological and epistemic grounds for assigning that meaning. Questions were asked about what a definition of a term really was; what to do when two different branches of biology used different terms to refer to the same phenomenon, or assigned different meanings to the same term; and what the relation was between a general definition, and empirical observations that did not seem to conform to that definition.

[2] This paper does not provide a detailed discussion of the history and epistemic role of GO, which can be found in publications by the GO consortium (2000, 2004, 2006, 2007), Lewis (2004), Smith et al (2007) and Leonelli (2008, 2009a/b, 2010 and in preparation).

[3] The order in which I approach the three perspectives and what can be learnt from them is arbitrary, as I hope will become clear when examining each viewpoint. I chose to start from philosophy because that was my own starting point in approaching this case.

After investigating the background for this meeting through scientific literature and conversations with the scientists involved, I was able to make sense of the participants' preoccupations with definitions. GO consists of a network of terms organised hierarchically through a simple set of relations (such as, for instance, 'the nucleus is *part of* the cell'). Each term is taken to refer to an existing biological entity or process (e.g. 'cell', 'root'), and it is used to classify datasets that are known to have evidential value with respect to that entity or process (e.g. data documenting the expression of genes involved in root development in Arabidopsis). Thus, GO terms can be used as database keywords to search for data of relevance to any biological phenomenon: researchers type in the term indicating the phenomenon that they wish to research, and in a matter of second they obtain a list of potentially relevant datasets taken from a variety of organisms. From a philosophical perspective, GO works as a *classification system* for datasets *and* biological phenomena: it defines phenomena in order to classify data (Leonelli 2009b).

Before the advent of GO, there was no way to conduct data searches that would cut across model organism databases – each database would only serve researchers working on a specific organism (Lewis 2004). Having a unique classification system for available data on the most popular model organisms was intended to radically increase the efficiency with which those data are shared across research communities, a crucial improvement in the context of cross-species, integrative research. As emphasised in the *Nature* paper that first presented the project to the biological community, the goal of GO was to "facilitate, in our writing as well as our thinking, the grand unification of biology that the genome sequences portend" (Gene Ontology Consortium 2000, 28).

According to this vision, unified access to data comes at the price of a unified classification system. The labels chosen for classification need to be intelligible to, and usable by, all potential users – a veritable 'lingua franca' (Gene Ontology Consortium 2000, 27). One way to achieve universal intelligibility is to elaborate explicit, precise, unique and unambiguous definitions for each term used. Rhee et al (2006, 345) point to the explicit and unambiguous quality of definitions in bio-ontologies as their most important characteristics:

> "The data that are generated and analysed as described in the previous sections need to be compared with the existing knowledge in the field in order to place the data in a biologically meaningful context and derive hypotheses. To do this efficiently, data and knowledge need to be described in explicit and unambiguous ways that must be comprehensible to both human beings and computer programs. An ontology is a set of vocabulary terms whose meanings and relations with other terms are explicitly stated and which are used to annotate data".

Access to definitions, available online as propositional statements, is supposed to enable users to check the meaning attributed to each GO term, and thus to dispel confusion about which phenomenon each term is meant to indicate. Hence, users should able to efficiently use GO terms to search for data.

Such reliance on language to disambiguate meaning raises a problem for the philosopher of biology, who is well aware that the interpretation of language depends on tacit knowledge and skills as much as it depends on convention - which means that individuals with different experiences might well attribute different meanings to the same statements.[4] Several philosophers have discussed the lack of consensus among biologists on how to define biological phenomena, and consequently on how to represent biological knowledge.[5] The vocabularies used differ depending on researchers' disciplinary affiliations, interests, skills and choice of model organisms; and this diversity is fully justified by the match between each specialised terminology and the materials, methods and tacit knowledge that that terminology has evolved to capture.

If constructing a unique, universal representation of biological knowledge were unproblematic, content meetings such as the one I attended in 2004 would not be necessary. Having to choose one definition in the presence of several alternatives, as in the case of 'pathogenesis', is what makes GO into something other than a neutral *lingua franca*. In response to this issue, GO curators have argued that terms in the classification system are not intended to substitute specialised terminology, but rather to complement it

[4] As emphasised by Claude Bernard already 150 years ago, 'when we create a word to characterise a phenomenon, we then agree in general on the idea that we wish it to express and the precise meaning we are giving to it; but with the later progress of science the meaning of the word changes for some people, while for others the word remains in the language with its original meaning. The result is often such discord that men using the same word express very different ideas' (Bernard 1927, 188).
[5] See Stotz et al 2004, and the collection Kellert et al 2006.

for the purposes of data dissemination and retrieval. Again, this distinction proves problematic: as highlighted by philosophers, the activities of contributing and accessing data, evaluating their significance, and conducting experiments are tightly interrelated.[6] Thus, it would be hardly possible for biologists to use different terminologies when conducting experiments and when curating data.

Seen from a philosophical perspective, the choice to rely on definitions to articulate the meaning of classificatory categories leads to one of two equally undesirable outcomes. The first is *inoperability* due to the mismatch between terminology used within GO and the vocabulary used by many of its users in their research practice. The second is a pernicious form of *terminological imperialism*. By settling on a specific definition, curators decide the meaning that all users of their tool should assign to the term in question. Users in need of data therefore accept the definitions assigned by curators and use them as substitute for their own local terminologies, with serious (and possibly negative) consequences for their research. Viewed in this way, GO becomes an exemplary case of knowledge unification through the top-down standardisation not only of the terms used, but also of the meaning assigned to them - hardly a new research strategy, as forming coalitions to impose one's own terminology (and related meanings) has been a standard move throughout the history of science.

This philosophical assessment certainly illuminates key tensions characterising the making of GO. However, when thinking along these lines in 2006, I was well aware that GO does not fit either of the two outcomes suggested above. It certainly is not inoperable, as increasing numbers of databases are using bio-ontologies to classify their data. And viewing it as an imperialistic exercise is in stark contrast with the curators' own description of their attempts as a "service" to the biological community (Lewis 2004). Far from being unaware that their project might be perceived as dominating and/or removed from experimental biology, the curators emphasise that GO was built precisely in order to respond to those dangers: "the GO ontology and annotations are continually updated to reflect current knowledge, to correct errors and to improve logical consistency. The GO ontology is updated daily and most of the annotation files are released weekly" (Rhee et al 2008).

[6] See for instance Duhem (1974) and Hacking (1983).

Without knowing precisely how GO was being developed, updated and critiqued by its users, it was hard for me to assess whether these claims were realistic and thus whether GO was managing to escape its imperialist fate as a classification system. Could it be that GO was different from other classification systems by virtue of how it was maintained? Philosophy was not going to help me answer that question. To better grasp the role of GO within contemporary biology, I would have to investigate the social structure through which GO is maintained. In practice, this involved interviewing GO curators, visiting the premises in which they work, attending numerous scientific meetings and interviewing users; in methodological terms, it implied using ethnographic research skills such as participant observation and discourse analysis of interview transcripts. From a disciplinary viewpoint, I was temporarily leaving philosophy to venture into the sociology of bioinformatics.

2. Sociological perspective: GO as an expert community

The first thing I realised when venturing into the sociological part of my research is that GO is not just a digital tool available on the internet for the easy retrieval of data. It is also the people responsible for maintaining this tool, who have actually developed a new kind of expertise for this purpose: the professional figure of the curator.

The original creators of GO, all of whom are distinguished experimental biologists, have always been aware of the changing and pluralistic nature of biological knowledge, and thus of the tension between the instability of such knowledge and the stabilising effect of representing it through a set of well-defined terms. Resolving this tension was the challenge that the curators set out to address in the first place. Remarkably, the solution did not come in the form of a specific structure for bio-ontologies as classification systems, but rather of the opportunity to modify that structure through human agency and technological innovation. The Internet was making it possible to treat GO as a tentative representation of biological knowledge, subject to constant modifications depending on its usefulness to experimental biologists. For this vision to come true, someone had to make the effort of encouraging critical feedback on GO and translating that feedback into appropriate changes to the system. Experimental biologists did not seem to be interested in that job, busy as they are in producing publishable and

patentable results. This left a professional niche open for bio-ontology curators, who saw themselves as taking over the responsibility (and associated efforts) for updating the classification system so that it would mirror as closely as possible the research practices of its users.

Thus the curators' strategy to avoid terminological imperialism consisted in restructuring the division of labour between experimenters and bioinformaticians, advertising themselves as "mediators" between different epistemic communities, and transforming database management into a scientific task requiring appropriate expertise (Leonelli 2010). This became clear to me while studying the practices through which curators maintain GO, and the ways in which this work requires their active judgement and intervention. A remarkable activity carried out by curators is, for instance, the *extraction* of data from scientific publications, so that they can be classified through bio-ontology terms. In the course of interviews, I realised that curators cannot compile data from all available publications on any specific gene product. Often there are too many publications, some of which are more up-to-date than others, some of which are more reliable than others because they come from more reliable sources. Curators need to be able to select one or two publications that can be 'representative' for any given gene product. They have to assess what data can be extracted from those papers, and whether the language used within them matches the terms and definitions already contained in the bio-ontology. Does the content of a paper warrant the classification of the data therein under a new bio-ontology term? Or can the contents of the publication be associated to one or more existing terms? These choices are unavoidable, yet they are impossible to regulate through fixed and objective standards. The reasons why the process of extraction requires manual curation are also why it cannot be divorced from subjective judgement: the choices involved are informed by a curator's ability to understand both the original context of publication and the context of bio-ontology classification.

Other examples of interventions by curators emerge when analysing data collection processes. An important criterion for data to be included in GO is that they are computationally manageable. Several types of data (such as pictures and tissue samples) are either unavailable in digital forms or available in a wide variety of formats, not all of which are compatible with the software and format supported by GO. In those cases, curators need to marshal data so that they fit the GO format. Further, there is the question of formulating definitions for GO terms. As I mentioned, these definitions provide an

unambiguous identification of the entities and processes in the real world to which the terms refer. At the same time, they are geared to fit as many organisms as possible, thus specifying features associated with the term that may or characterise only a subset of existing species. They also need to accommodate critiques provided by experimental biologists, no matter their discipline or training. Constructing a definition that fulfils all these criteria constitutes a huge conceptual challenge. Last but not least, collecting data and linking them to specific terms is not enough for users to be able to re-use the data. To assess the evidential value of data, users need to know what organism those data were taken from; which laboratory produced them, and for which purpose; and which instruments and protocols were used. This information, dubbed "meta-data", is what enables users to interpret the biological significance of data; yet, again, there is no objective standard to determine which bits of information about an experiment should count as meta-data.[7] Curators are thus also responsible for choosing which bits of information about the provenance of data users should be able to access through GO.

Through the above processes, curators mediate between the requirements of their classification system (consistency, computability, ease of use and wide intelligibility) and the results, assumptions and practices characterising the work of bio-ontology users. They have developed a specific expertise enabling them to do this, encompassing a good understanding of programming, decent IT skills, a basic training in various biological disciplines and familiarity with experimentation at the bench (so that they can understand specific experimental settings and anticipate the needs of the users of bio-ontologies). Further, they have developed social networks enabling them to consult and police each other. Immediately after the creation of the GO classification system, its curators founded the GO consortium, an organisation grouping together the scientists responsible for adding GO entries. One of the purposes of the consortium was to ensure that all curators were trained in similar ways, thus sanctioning the difference between the expertise of curators and other scientific expertises (Leonelli 2009a).

Understanding the practices and expertise through which GO is maintained is absolutely essential to understanding its role in contemporary biology. A sociological study of the GO community makes clear that the success of this bioinformatic tool is due as much to the way in which it is structured as to

[7] Those standards are being developed, but are yet far from achieving widespread recognition within the biological community (Taylor et al 2008).

the way in which it is maintained: GO is at once an epistemic and a social object. Curators themselves emphasise the social dimensions of this project as one of its main features. One of the GO creators, Suzanna Lewis, went so far as to define the collaborative network built around GO as 'the single largest impact and achievement of the Gene Ontology' (Lewis 2004, p. 103). And indeed, GO curators have consistently placed 'community involvement' as the factor that most contributed to their success (Gene Ontology Consortium 2004; Bada et al 2004).

Such emphasis on the collaborative dimensions of GO is understandable in the light of the efforts put by curators into establishing feedback mechanisms from users. However, my interviews with users and curators yielded a finding that threatens to upset the emerging picture of GO as a collaborative project: the relations between curators and users remain uneasy. All the curators I interviewed emphasised the difficulties they encountered in eliciting feedback from GO users. Interviews with users corroborated this finding. Users have little time to familiarise themselves with GO, and thus to engage in providing critical feedback: they wish to get data out of databases and go back to experimental work. Given these constraints and interests, users are happy to accept the curators' authority on how data are to be classified and distributed. According to several interviewees, it is the curators' responsibility to provide a system that matches the needs of experimental research, and it is not the users' job to tell them how to do this. The result is a vicious circle: to gain users' trust, curators needed to successfully establish themselves as epistemic authorities; yet such authority ends up undermining their dialogue with users, with damaging consequences for GO itself. Without user feedback, GO cannot claim to build on 'community involvement'; nor can it claim to elude the threat of terminological imperialism.

Having realised the extent of the problem, curators have been attempting to develop solutions to it over the last four years. The result has been the development of a third dimension of GO: that of a growing scientific institution playing a crucial role in the governance of data sharing practices. To address this issue, I had to investigate the ways in which GO curators have institutionalised their relationship with users, as well as with the economies of vitality in which research is situated. This meant approaching my case from the perspective of contemporary history, exploring both the roots and development of GO over time, and the ways in which this initiative fit the wider shifts in science policy characterising the last two decades.

3. Historical perspective: GO as an emerging institution

When adopting an historical perspective, attention is immediately drawn to the ways in which the practices and methods characterising GO have become increasingly more articulated and centralised since its creation. This was both a result and a cause of the increase in size and impact of the project itself. In less than a decade, what used to be a haphazard list of biological terms has become one of the most complete and polished classification systems in current molecular biology; and what used to be a group of like-minded colleagues committed to the same goal – that of cross-species data integration – has become a managerial committee responsible for key aspects of the governance of data sharing in biomedicine. These include the establishment of *rules* specifying what counts as 'good practice' in annotation; how bio-ontologies should be constructed; and how users should contribute to the development of annotations. An example is the true path rule, which states that "the pathway from a child term all the way up to its top-level parent(s) must always be true" (GO Website, accessed 25/03/2009). This means that, every time a new term or gene product is added to GO, curators need to check that its connection to existing terms holds true. This rule is key to updating GO, since adding terms to the network often has repercussions on the significance of the whole structure, rather than simply on the terms directly connected with the new entries.

To make sure that its curators would gain the visibility and power necessary to enforce those rules, the institutional incarnations of GO have multiplied. For instance, the Open Biomedical Ontologies consortium was created in 2005, under the auspices of the National Centre for Biomedical Ontology, to group together all bio-ontologies that are constructed similarly to GO (Rubin et al, 2006). At the same time, a smaller committee within the OBO consortium, including the original creators of GO, was assembled under the name of OBO Foundry to police the development and use of rules for making bio-ontologies. Further, several GO members were involved in the creation of Biocurator.org, a network helping database curators to communicate with each other and establish general norms of good practice (http://www.biocurator.org; Howe et al 2008). Through the institution of these kinds of networks, the

GO community has become a regulatory centre for data sharing practices – what I call a *labelling centre* (Leonelli 2009a).

Crucial to explaining the explosive development and success of GO as a scientific institution are its timeliness and fit with the broader political and scientific context. A fuller understanding of GO as a bioinformatic tool needs to consider its ties to the globalised institutional arrangements characteristic of 21st century science. Historians and sociologists are studying the increasing emphasis on collaboration and interdisciplinarity characterising the last twenty years of science policy.[8] Amidst rapid shifts in communication technologies and international relations, funding bodies in the States, the European Union and Japan have strongly supported cyberinfrastructure as a way to cut costs research costs (Buetow 2005); to enhance interdisciplinary 'collaboratories' (Finholt 2002); and to revolutionise research through reliance on data-driven methods (e.g. genome-wide association studies, microarray diagnostics, etc.; see Bell et al 2009).

GO fully embraced this turn towards collaborative work. As its scope and impact grew, the curators put increasing emphasis on the flexibility of the system, and its capacity to learn from its own mistakes by being constantly measured against experimental results (Hill et al 2008, 9). This emphasis was incorporated into the main tenets of the Open Biomedical Ontologies consortium (Smith et al 2007). Increasing attention was also devoted to possible misuse or misunderstandings by users. Once GO had established its high status as a bioinformatic tool, curators could afford to reproach the users for their reluctance to engage with GO:

> Although GO is a powerful tool, researchers who use it should be cognizant of the features of the ontologies and annotations to avoid common pitfalls. Available annotation for a given organism might affect results and conclusions. Therefore, care should be taken when choosing an analysis method; it might be essential to include or exclude certain types of annotations for certain types of analysis. In addition, it is crucial for any analysis to cite data sources (including the version of ontology, date of annotation files, numbers and types of annotations used, versions and parameters

[8] Exemplary for this work are the collections edited by Fortun and Mendelsohn (1999), Gaudillière and Rheinberger (2004) and Atkinson et al (2009).

> of software, and so on) to ensure that results are fully reproducible. The GO is a tool that will become increasingly powerful for data analysis and functional predictions as the ontologies and annotations continue to evolve. Our hope is that researchers fully understand and thus can take full advantage of this vital resource (Rhee et al 2008).[9]

The curators have not limited their hopes to words. Rather, they are actively seeking to transform experimental research so as to make it aware of, and compatible with, bioinformatics. For instance, they are pushing journal editors to establish new rules for the submission of research papers, thanks to which papers would not be accepted unless they come with GO annotations. If successful, this change in submission requirements will effectively force experimenters to critically engage with GO; the curators hope that this will spur an improved and better informed dialogue between biologists and bioinformaticians (at the same time, experimenters might of course perceive this as a sign of imperialistic aspirations). Another powerful weapon used by curators is education. Several database curators with GO affiliations have promoted the use of GO through workshops at conferences and research institutes. Thanks also to their campaigning, basic training in experimental biology (at least in the US, UK and Germany) now often incorporates bioinformatics.

Seeing the GO project through a historical perspective enables a more balanced answer to the challenge set by my philosophical reading of this tool. GO is imperialistic in the sense that it is becoming institutionalised as a new standard for data sharing. Its success as a classification tool implies that it is becoming a powerful component of the new "communication regime" imposed by the development of cyberinfrastructure as a whole (Hilgartner 1997). Like all social and epistemic entities, GO needs constant work and support to survive. In other words, it needs social and economic power, which can only come from enforcing its use by the biological community at large, thus becoming an indispensible tool for biological research. This domineering attitude does not necessarily imply that GO is a static classification system, imposing a fixed language to experimental biologists of all trades. As shown through my sociological analysis, the flexibility of GO and the care taken by curators to build in feedback are its key strengths. In fact, dialogue between users and curators is made possible by the very mechanisms that "force" users to adopt GO.

[9] For attempts to involve, train and/or police GO users, see also Hill et al 2008 and Howe et al 2008.

The Ontology Issue

In which sense is the Gene Ontology an ontology in the philosophical sense?[10] Not surprisingly, this is the most popular question asked whenever I present this research. I wish to consider it here as a further illustration of the analytic usefulness of the multidisciplinary approach that I have been arguing for. When thinking of GO from a philosophical viewpoint, I was first tempted to reply that this tool is simply not an ontology in the philosophical sense of the term. It expresses knowledge held by members of the biological community: what they know about biological entities and processes today, which differs from what they might know tomorrow. That fallibility makes it into an epistemic, rather than an ontological tool denoting what exists. What I knew of the history of the term 'ontology' in informatics seemed consistent with this assessment. The term is used in computer science to indicate a set of representational primitives with which to model a domain of knowledge or discourse. It was originally picked up from philosophy, but it quickly acquired a technical meaning that had everything to do with the structure of 'object-directed' programming languages and little to do with the intricacies of metaphysics (Gruber 1993). According to some of the curators I interviewed, it was this technical meaning that the bio-ontologies had taken up.

It is indisputable that the term ontology used in bio-informatics has robust links to computer science. Yet, no direct reference to information technology was made in the pioneer GO paper published in 2000. Early discussions of bio-ontology development, as recorded in the GO archives, were also focusing more on the underlying biology than on the programming language and structure needed to implement it – in any case, both elements were playing a key role. Further, several interviewees insisted that bio-ontologies have an ontological value in the philosophical sense – an insistence echoed in several publications which introduce the topic in ways similar to this: "ontologies have long been used in an attempt to describe all entities within an area of reality and all relationships between those entities" (Gene Ontology Consortium 2000, 27).

[10] The study of the nature of being, existence or reality and of the categories of being and their relations.

To make sense of these findings, I tried to elaborate a more sophisticated philosophical reading of what ontology might mean in the case of GO. As I mentioned, curators wish GO to capture the background knowledge that biologists take for granted when carrying out experiments. GO terms refer to (what biologists know about) actual entities and processes. GO could therefore be interpreted as a map of the ontological commitments underlying biological research – something close to what Lorraine Daston calls 'dynamic ontology created and sustained by scientific observation' (Daston 2008, 97). This intuition is corroborated by the observation that phenomena that do not get labelled with a bio-ontology term are destined to remain invisible to anyone using bio-ontologies to trace data and materials for their research. As argued by Geoffrey Bowker (2006), and often debated in philosophical discussions surrounding the notion of natural kinds[11], any classification inevitably defines an ontology, and there are consequences to what is excluded. Interpreted in this way, GO does constitute an ontology in a philosophical sense, even if not in the Platonic sense of defining the immutable essence of biological phenomena (Leonelli in preparation).

Thus, the joint use of philosophical and sociological material has given me one useful way to think about the ontology issue. Historical analysis has given me another. Last year I started to question my interviewees about the history of the term ontology in their work, rather than about the meaning that they thought it held now. In other words, I shifted my interviews from a social science perspective, where the analyst tries to capture his subjects' thoughts and their relation to their practices, to an oral history perspective, where subjects are asked about their perception of their role in a specific event (in this case, the creation of GO). This shift yielded an interesting result. A major figure in the development of GO argued that the original choice of the term 'ontology' did not stem from deep philosophical concern, nor from an attempt to mimic the vocabulary used in computer science. Rather, my interviewee thought of it as parodying the ways in which computer scientists, and particularly researchers in medical informatics, used the term ontology. This remark led me to think about the notoriously problematic relation between biology and information technology (Goble and Wroe 2004) as very relevant to explaining the meaning currently acquired by the term ontology in bio-informatics.

[11] Dupré (1993), Reydon (2009).

Especially at the beginning of the project, there was abundant hostility between GO curators and information scientists on how to think about data sharing. One of the key motivations fuelling the birth of GO was to provide an alternative to classificatory ontologies built purely on the basis of programming rules. The argument was that the point of distributing genomic data through databases was to enable biologists to re-use them for their own purposes. A viable classification system has to make biological sense: classifying data for dissemination requires a deep awareness of how biologists might re-use those data, which informs a view of how databases need to be structured. The creators of GO were adamant that information technology alone could not yield useful databases. It had to be put at the service of contemporary biology. Seen in this context, the name Gene Ontology was indeed chosen as a provocation, an attempt to create something other than the ontologies developed by computer scientists. This insight allowed me to see the ontology issue under a new light. From a philosophical perspective, I could still develop ideas about its relation to traditional views on ontology. From a historical perspective, however, this term had neither a philosophical nor an informational meaning: it is used by GO curators to define a classificatory structure that they perceive as innovative.

Another historical insight that helped my thinking about this issue was the fact that a philosopher played a prominent role in developing GO. In the early 2000s metaphysician Barry Smith, who had written on the relation between his approach to metaphysics and the notion of ontology in computer science, published a series of critiques of GO, arguing that the curators' way of constructing ontologies was philosophically problematic. Interested in what they could learn from his approach, the curators welcomed Smith into their midst, ultimately enabling him to become the Director of the National Institute for Biomedical Ontologies and the main architect of the OBO Foundry. Smith used his background in philosophy and computer science to claim expertise on the term ontology. In so doing, he effectively brought a specific philosophical tradition to bear on the original, non-philosophical meaning assigned to the term by bioinformaticians. His work can thus be seen as one of the causes for the interdisciplinary mingling that is responsible for the 'mixed status' of the term ontology in GO. Viewing Smith's role in an historical perspective helps to make further sense of the current conflation between philosophical, informational and biological notions of ontology.

Conclusion: Integrating HPSSB

Using GO as a case study, I have been arguing for the need to study bioinformatics through multiple disciplinary lenses. A philosophical approach is indispensible to understand the epistemic significance of and the intrinsic problems with constructing a system such as GO. A sociological analysis reveals how GO works and what makes it potentially different from other classification system, namely the fact that its development is in the hands of an expert community focused on dialogue with the broader biological community. And a historical perspective reveals the strong tie between the features of GO and the globalised institutional arrangements characteristics of 21st century science. In the process of developing GO, its curators have become accountable to the wider scientific community for their role in governing data dissemination.

I conclude that GO is best understood as encompassing three entities at the same time, each of which plays a crucial role in defining the significance of this tool for contemporary biology: it is a *classification system* created to facilitate the distribution and use of data as evidence for new insights; an *expert community* specialised in the curation of genomic data; and a *scientific institution* regulating data sharing practices. These three dimensions of GO can be clearly distinguished analytically, but are tightly intertwined in practice. I wish to suggest that all bioinformatic tools are at once epistemic, social and institutional entities, since they shape the biological knowledge extracted from data and at the same time regulate the organisation, development and communication of biological research. Only the integrated use of history, philosophy and social studies of biology (commonly referred to as HPSSB) makes it possible to capture such complex, multifaceted and possibly more truthful analysis of what bioinformatic tools are and mean for contemporary biology.

Like any research result, the picture of GO painted through HPSSB is by no means (nor it can ever be) complete, yet appealing to these different perspectives rescues the analysis from some potential pitfalls. If stuck solely with a historical perspective, awareness of the epistemic value of GO as a classification system today might be lost, together with knowledge of the kind of expertise used by the curators to maintain this system. Reliance on philosophy alone might be equally misleading, pushing the analyst to think of GO as an imperialistic attempt to take over biological terminology – a reading that again

captures one important dimension, but which overlooks the consortium's attempts to avoid such a domineering attitude and to structure the ontology in dialogue with its users. A sociological analysis is indispensible for understanding the complex interface between curators and users of GO, as well as its broader scientific, political and economic context. Yet, it does not illuminate the intricacies and implications of constructing a classification system focused on the propositional definition of phenomena, nor does it encourage an assessment of how GO was born and how it has developed over the years – which, as we have seen, is crucial to grasping the paradoxical nature of this system as a regulatory institution as well as a service to biologists.

Underlying my view are two broad methodological claims. The first is that history and sociology are complementary from both an intellectual and a methodological viewpoint. Intellectually, it is well-accepted to seek continuity between social and historical understandings, certainly when taking seriously the Foucauldian invitation towards a genealogical approach to contemporary social phenomena. The methodology required to follow this through, however, lags behind. Historians of the 20th century make abundant use of oral history, without necessarily learning from interviewing techniques used in the social sciences, where questions about reflexivity, use of transcripts and the role of the interviewer in extracting information are matters of daily concern. Conversely, social scientists can learn from the historians' attention to archival sources, and the care with which they are positioned within specific spatio-temporal contexts in order to be interpreted. The second methodological claim is that, just as much as history and sociology, philosophy of science requires an empirical footing. This claim is corroborated by several movements within the discipline today, ranging from 'experimental philosophy' to the 'philosophy of science in practice'. Nevertheless, there is little understanding of which methods philosophers should adopt to provide evidence for their claims. One possibility is to exploit historical methods to analyse archives documenting past scientific practices. Another way is to conduct ethnographic studies of scientific practice, including visits to sites of scientific research and interviews to actors. As illustrated by my struggle to tackle the GO ontology question, more work needs to be done on how these different methods can be combined within the same study.

Among the various possible objections to an integrated HPSSB approach, there are two major practical concerns I wish to mention in closing. One is the pedagogical implications: the need to re-think

radically how scholars in these fields need to be trained to conduct high-level research on scientific practices. The other is the sheer scale of the research effort involved, as the time-scale, skills and complexity involved in conducting a study of this kind are far beyond what a single individual could achieve, especially given the short-term requirements for 'results' of today's academia. Both of these concerns seem to me simply to indicate the shifting state of affairs in the social sciences and humanities. One radical lesson learnt from studying contemporary collaborative networks in biology is that no set of sciences, including the social sciences, can afford a narrow, disciplinary perspective on phenomena. I certainly could never get my head around the issues arising from my multi-disciplinary approach without the help of specialists within each disciplinary community, which implies constant dialogue with several different scholarly networks as well as the need to accept that one's results can be overturned at any point by insights never considered before. Science studies focused on bioinformatics require extensive collaborative networks, in the same way as the natural sciences do when looking at biological phenomena. Studying bioinformatic tools requires the integration of some expertise in computer science, biology, philosophy, history and sociology – plus, arguably, economics and anthropology, which I did not discuss here. This shift towards collaboration is evident in the number of societies, journals and centres devoted to HPSSB around the globe. Still, it does not yet fit the institutions, publishing venues and training programmes characterising mainstream social science and humanities, which still favour single disciplinary perspectives and individual modes of investigation over collaborative projects.

Acknowledgments

This research was funded by the UK Economic and Social Research Council as part of the ESRC Centre for Genomics in Society (University of Exeter). I am grateful to Vincent Ramillon and Edna Suarez for inviting me to contribute to their excellent workshop and to this publication; and to my colleagues at Egenis, especially Maureen O'Malley, Barry Barnes and Hannah Farrimond, for helpful discussions.

Bibliography

Atkinson P., Glasner P. and Lock M. (eds.) *The Handbook for Genetics and Society: Mapping the New Genomic Era*, London: Routledge.

Bada M. et al, 2004, "A short study on the success of the Gene Ontology", *J Web Semant*, 1: 235-240.

Bell G., Hey T. and Szalay A., 2009, "Beyond the data deluge", *Science* 323: 1297-1298.

Bernard C, 1927 [1855], *An Introduction to the Study of Experimental Medicine*, Dover Publications.

Bowker G.C., 2006, *Memory Practices in the Life Sciences*, The MIT Press.

Buetow K.H., 2005, "Cyberinfrastructure: Empowering a 'third way' in biomedical research", *Science* 308, 5723: 821 – 824.

Daston L., 2008, "On scientific observation", *Iris* 99: 97-110.

Duhem P., 1974 [1914], *The Aim and Structure of Physical Theory*, New York: Atheneum.

Dupré J., 1993, *The Disorder of Things*, Harvard University Press.

Finholt T.A., 2002, "Collaboratories". In: Cronin B. (ed), *Annual Review of Information Science and Technology*, American Society for Information Science and Technology, 36: 73-107.

Fortun M. and Mendelsohn E. (eds.), 1999, *The Practices of Human Genetics*, Springer.

Gaudillière J. and Rheinberger H., 2004, *From Molecular Genetics to Genomics*, Routledge.

Gene Ontology Consortium, 2000, "Gene Ontology: tool for the unification of biology", *Nature Reviews: Genetics*, 25: 25-29.

Gene Ontology Consortium, 2004, “The Gene Ontology (GO) database and informatics resource”, *Nucleic Acids Research*, 32: D258-D261.

Gene Ontology Consortium, 2006, “The Gene Ontology (GO) project in 2006”, *Nucleic Acids Research*, 34: D322-D326.

Gene Ontology Consortium, 2007, “The Gene Ontology (GO) project in 2008”, *Nucleic Acids Research*, 36: D440-444.

Goble C. and Wroe C., 2004, “A tale of two households”, *Comp Funct Genom* 5: 623–632.

Gruber T.R., 1993, “A translation approach to portable ontology specifications”, *Knowledge Acquisition*, 5, 2:199-220.

Hacking I., 1983, *Representing and Intervening*, Cambridge University Press.

Hilgartner S., 1997, "Biomolecular databases: New communication regimes for biology?", *Science Communication* 17, 2: 506-22.

Hill D.P., Smith B., McAndrews-Hill M.S. and Blake J.A., 2008, “Gene Ontology annotations: what they mean and where they come from”, *BMC Bioinformatics* 9, Suppl 5:S2.

Howe D., Rhee S. et al, 2008, “The future of biocuration”, *Nature,* 455, 4: 47-50.

Kellert S.H., Longino H.E. and Waters C.K. (eds.), 2006, *Scientific Pluralism*, University of Minnesota Press.

Leonelli, S., 2008, “Bio-ontologies as tools for integration in biology”, *Biological Theory* 3, 1: 8-11.

Leonelli S., 2009a, "Centralising Labels to Distribute Data: The Regulatory Role of Genomic Consortia". In: Atkinson P., Glasner P. and Lock M. (eds.) *The Handbook for Genetics and Society: Mapping the New Genomic Era*, London: Routledge, pp. 469-485.

Leonelli S., 2009b, "On the locality of data and claims about phenomena", *Philosophy of Science* 76, 5.

Leonelli S., 2010, "Packaging data for re-use: Databases in model organism biology". In: Howlett P. and Morgan M.S. (eds.), *How Well Do "Facts" Travel*, Cambridge University Press.

Leonelli S., in preparation, "On the role of theory in data-driven research: The case of bio-ontologies".

Lewis S.E., 2004, "Gene Ontology: Looking backwards and forwards", *Genome Biology*, 6, 1: 103.

Reydon T.A.C., 2009, "Natural kind theory as a tool for philosophers of science". In: Dorato M., Rédei M. & Suárez M. (eds), *Proceedings of the first conference of the European Philosophy of Science Association*, Dordrecht: Springer.

Rhee, S.Y., Dickerson, J. and Xu, D., 2006, "Bioinformatics and Its Applications in Plant Biology", *Annual Review of Plant Biology*, 57: 335-360

Rhee S.Y., Wood V., Dolinski K., Draghici S., 2008, "Use and misuse of the gene ontology annotations", *Nature Reviews Genetics*, 9: 509-515.

Rose N., 2007, *Governing the Present*, Polity Press.

Rubin D.L. et al, 2006, "National centre for biomedical ontology: Advancing biomedicine through structured organisation of scientific knowledge", *OMICS* 10, 2: 185-198.

Smith B. et al, 2007, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", *Nature Biotechnology* 25, 11: 1251-1255.

Stotz K., Griffiths P.E. et al, 2004, “How scientists conceptualise genes: An empirical study”, *Studies in History & Philosophy of Biological and Biomedical Sciences,* 35, 4: 647-673.

Suàrez-Diàz E. and Anaya-Muñoz V., 2008, “History, objectivity, and the construction of molecular phylogenies”, *Stud. Hist. Phil. Biol. & Biomed. Sci.* 39: 451-468.

Taylor C. et al, 2008, “Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project”, *Nature Biotechnology* 26, 8: 889-896.